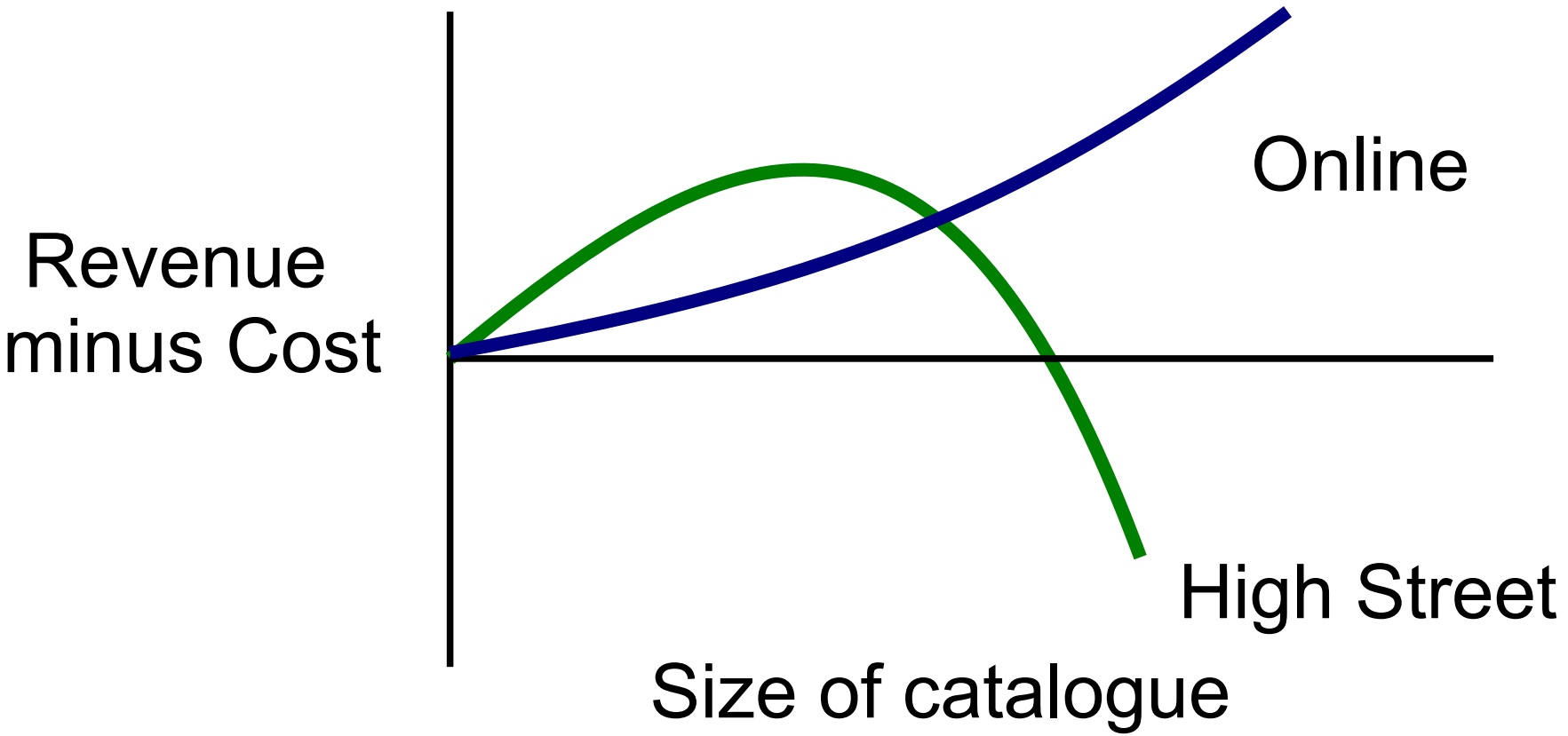
Implicit Data - a personal project

or

Recommender Systems
and
Scottish Country Dancing

David McQuillan

Why do they do it?

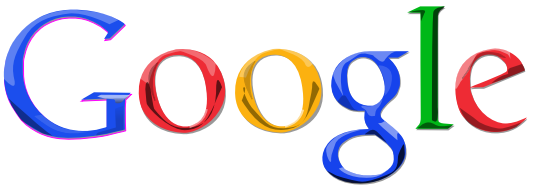


Want customers to spend money,
not spend time browsing the catalogue

What is a recommender system

As it says on the tin – they recommend

The system rates items based on an assessment of how well they match your preferences

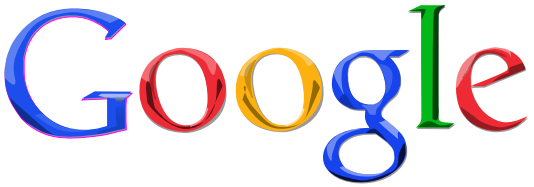


Explicit and Implicit

- Early system Tapestry

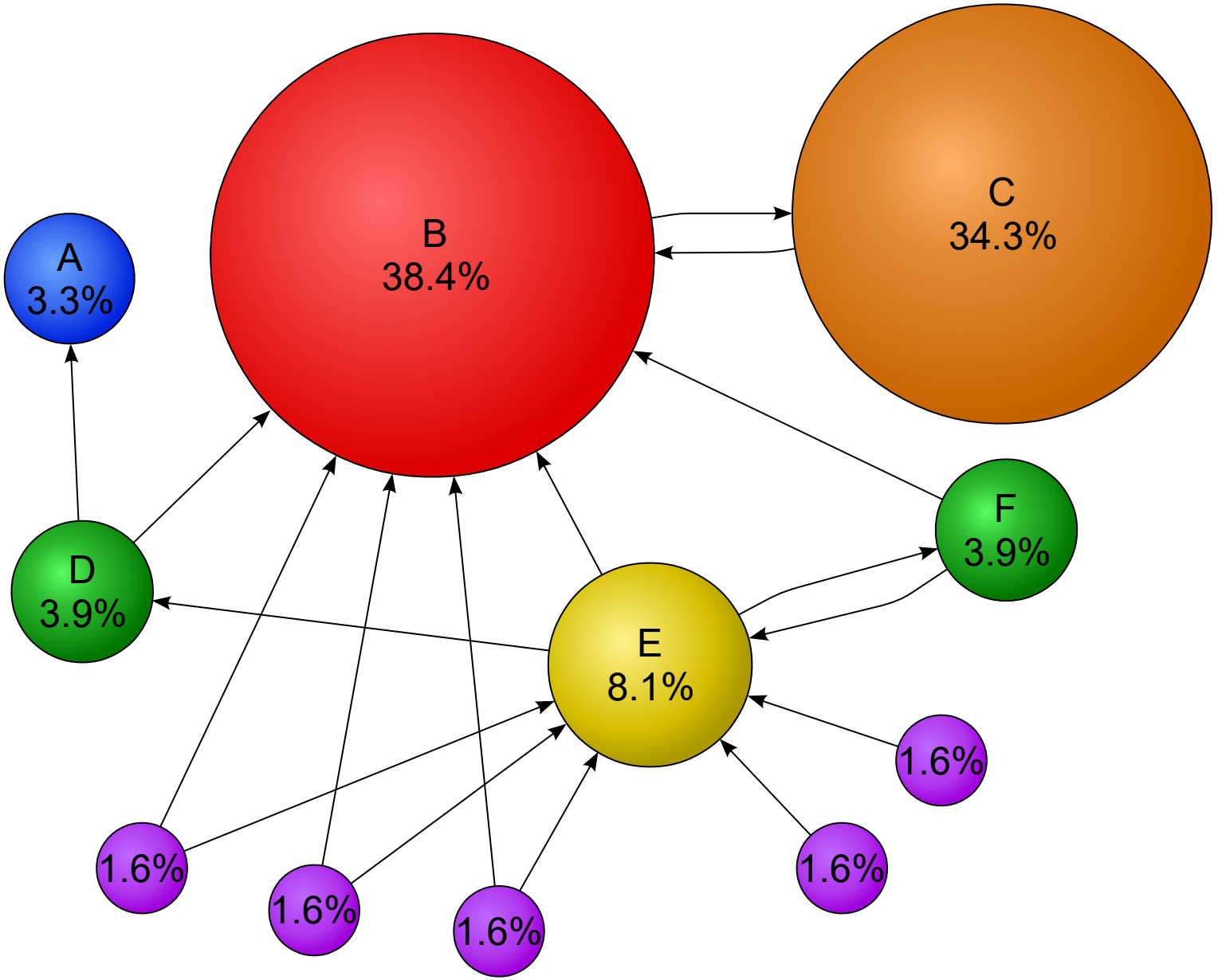
  Collaborative message filtering system
  Readers annotated messages with tags
  Can filters on tags and who set the
- Bobby can get messages Amy said were funny
- Explicit system
- How does one set up a good filter?

Implicit system



- Google Pagerank. No user profile
- The probability that a random link clicked will end at a page – assuming the current page is selected according to Pagerank.

$$\mathrm{PageRank}(u) = \frac{1 - \mathrm{dampening}}{\mathrm{NumPages}} + \mathrm{dampening} \times \sum_{v \in \mathrm{LinksTo}(u)} \frac{\mathrm{PageRank}(v)}{\mathrm{LinkCount}(v)}$$


A
3.3%
B
38.4%
C
34.3%
D
3.9%
E
8.1%
F
3.9%
1.6%
1.6%
1.6%
1.6%
1.6%

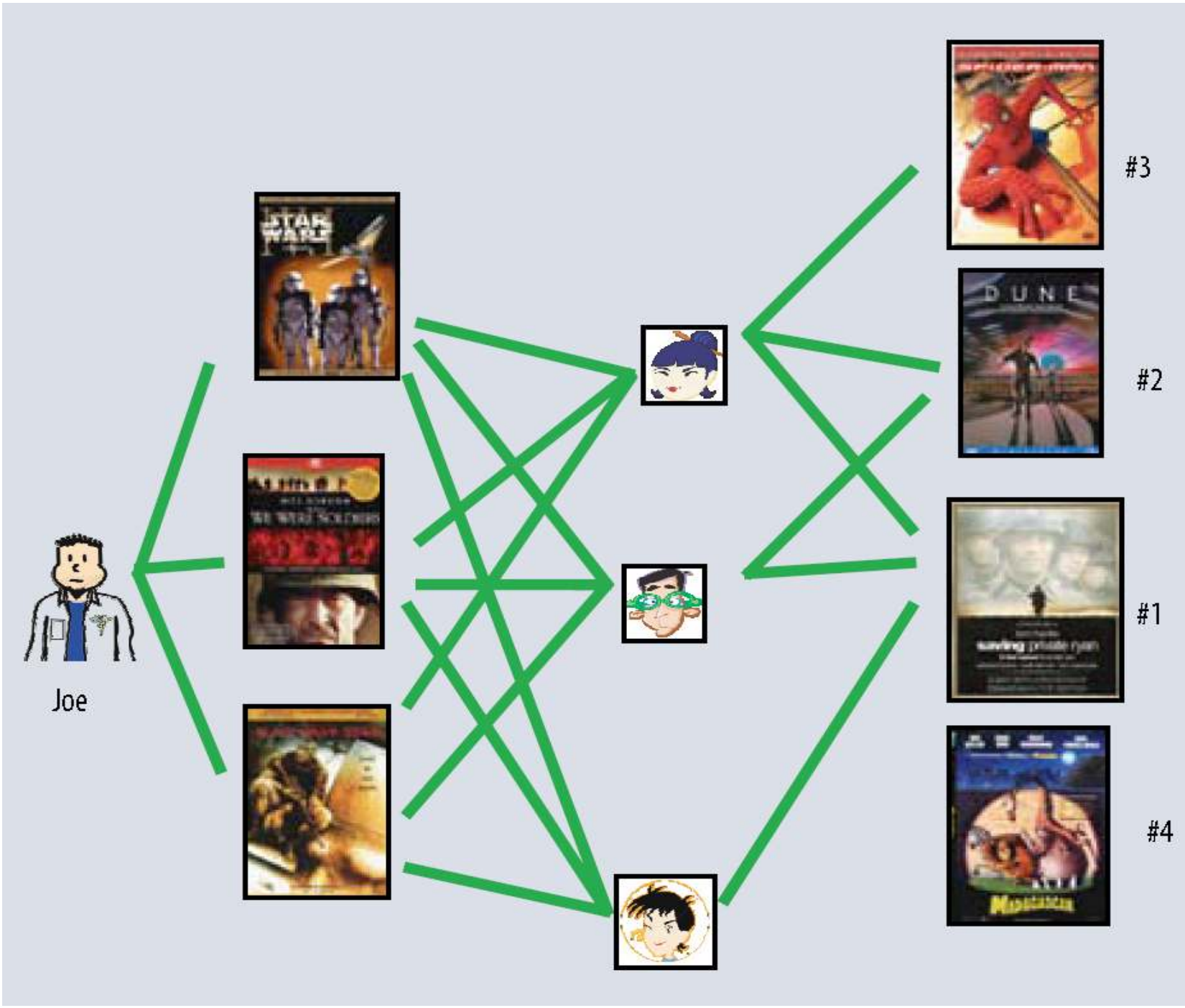
Recommender

amazon.com

- Amazon matched users to other users with similar purchases
- Early system. Users a, b vectors of N items

  find K of highest similarity $$S(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

  rating $$r(\mathbf{a})[i] = \frac{\sum_{\mathbf{b}\in K(\mathbf{a})} S(\mathbf{a},\mathbf{b})\cdot r(\mathbf{b})[i]}{\sum_{\mathbf{b}\in K(\mathbf{a})} S(\mathbf{a},\mathbf{b})}$$


Joe
#3
#2
#1
#4

What's implementation like?

- 400,000 orders a day at Amazon and probably over 10 million items for sale
- Naive implementation would need to do over 50 million similarity scores a second
- Even more factors actually taken into account

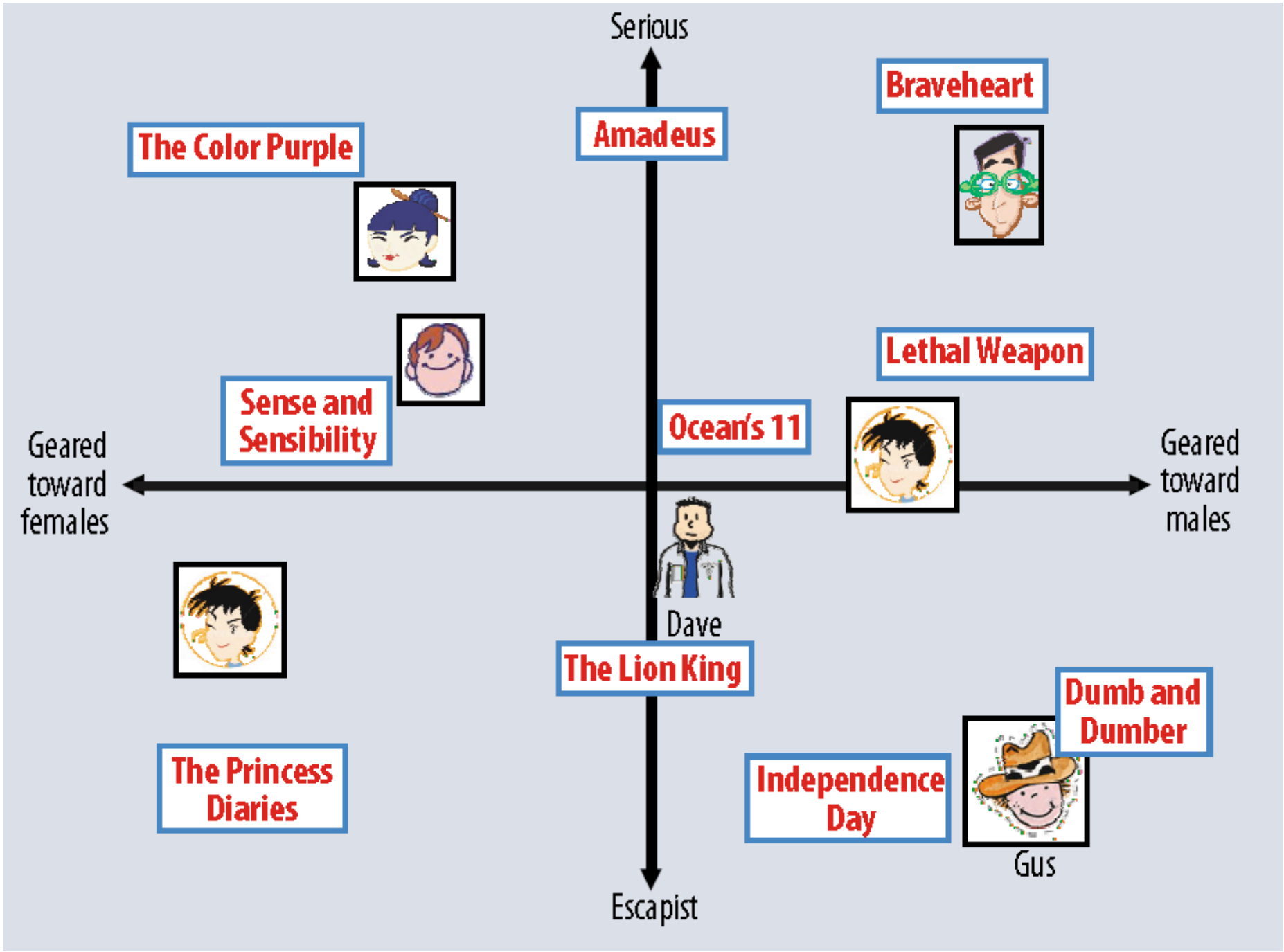
→ Item based precalculated similarity ratings

- However even a well tuned version of a recommendations system requires a enormous amount of computer power.

Netflix Prize

- Database of 100 million entries in time order

  (user id, movie id, rating)
- $1 million prize
- Target 10% smaller variation from the real ratings on a separate check database than an algorithm Netflix used. Their algorithm was rather like the Amazon one.


Serious
Escapist
Geared toward females
Geared toward males
Amadeus
Braveheart
The Color Purple
Sense and Sensibility
Lethal Weapon
Ocean's 11
Dave
The Lion King
The Princess Diaries
Dumb and Dumber
Independence Day
Gus

Singular Value Decomposition

- Any m×n matrix $\mathbf{M}$ can be factorized as

$$\mathbf{M} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^*$$

- Express an array as a sum of outer products with the first products making the largest contribution.

$$\mathbf{M} = \sum_i \sigma_i \mathbf{U}_i \otimes \mathbf{V}_i^*$$

Sum of Outer Products

$$\begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 0.591 & -0.807 \\ 0.807 & 0.591 \end{bmatrix} \times \begin{bmatrix} 6.318 & 0 & 0 \\ 0 & 0.274 & 0 \end{bmatrix} \times$$

$$\begin{bmatrix} 0.221 & -0.786 & -0.577 \\ 0.570 & 0.585 & -0.577 \\ 0.791 & -0.201 & 0.577 \end{bmatrix}^{\mathrm{T}}$$

$$= \begin{bmatrix} 0.826 & 2.129 & 2.955 \\ 1.127 & 2.905 & 4.035 \end{bmatrix} + \begin{bmatrix} 0.174 & -0.129 & 0.045 \\ -0.127 & 0.095 & -0.033 \end{bmatrix}$$

Unknown entries

Ratings $\mathbf{r}$ Item factors $\mathbf{q}$ User factors $\mathbf{p}$

$$\mathbf{r}[u,i] = \mathbf{r}_{ui} = \mathbf{q}_i^* \cdot \mathbf{p}_u = \sum_f \mathbf{q}^*[i][f] \cdot \mathbf{p}[u][f]$$

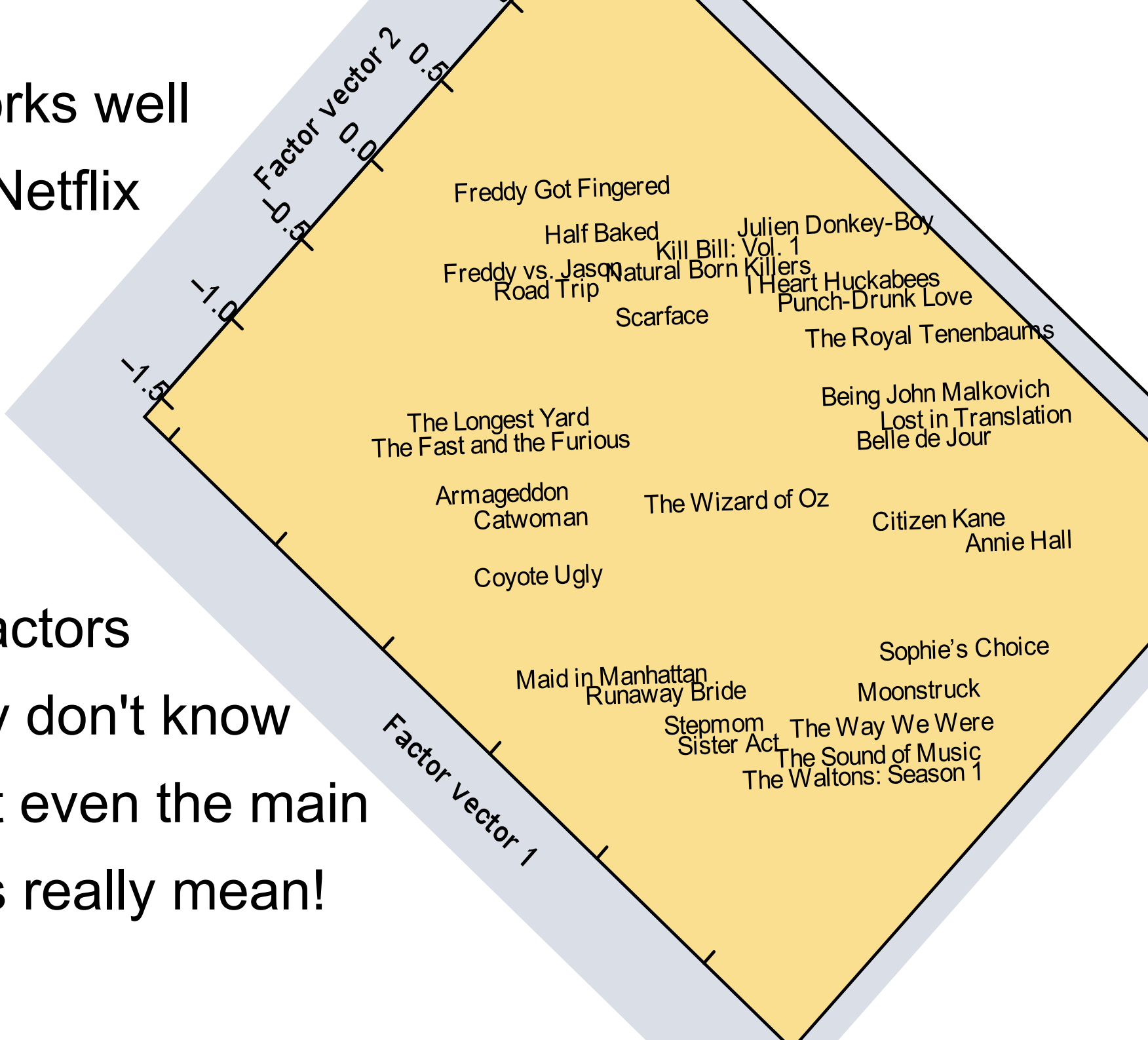
$$\min_{q,p} \sum_{\exists \mathbf{r}_{ui}} (\mathbf{r}_{ui} - \mathbf{q}_i^* \cdot \mathbf{p}_u)^2 + \lambda(|\mathbf{q}_i|^2 + |\mathbf{p}_u|^2)$$

In practice need an incremental learning system

plus some heuristics catering for instance for biases.

It works well

For Netflix



50 factors

They don't know what even the main ones really mean!

FLEET SCOTTISH COUNTRY DANCE SOCIETY

TREGOLLS OPEN AIR SUMMER GARDEN PARTY DANCE

To be held at

No.1 Tregolls Drive off Avenue Road, Farnborough GU14 7BN

(With thanks to Frank Bisby for the use of his garden and home)

Saturday 9 June 2012

2.00pm - 4.30pm followed by buffet table and tea/coffee

TICKETS £4.00 - AT THE DOOR

M/CS – Various

1	A Trip to Bavaria	Reel	* Leaflet (8 x 32)
2	Haste to the Wedding	Jig	RSCDS Bk 25/6 (8x32)
3	Sugar Candie	Strathspey	RSCDS Bk 26/9 (8x32)
4	Clutha	Reel	* RSCDS Bk 31/2 (SQ, 4x48)
5	Alison Rose	Strathspey	Imperial Bk 2 (4x32)
6	The Last of the Lairds	Jig	RSCDS Bk 22/5 (8x32)
7	The Jubilee Quadrille	Reel	Leaflet - Alan Macpherson (SQ set, 88 bars)
8	John McAlpin	Strathspey	Foss Galloway Album (8x32)
9	Seton's Ceilidh Band	Jig	* Morison's Bush (4x64)

INTERVAL (approx 3.15 to 3.30 FOR REST, IRN BRU etc)

10	The Duke and Duchess of Edinburgh	Reel	* RSCDS Bk 39/7 (8x40)

Dances in decreasing frequency order in South East

Wider is the South East. Local is the BHS branch.

Dance	Description	Hard	Local Frequency	Wider Frequency
Mairi's Wedding	**R8x40 3C/4**	2.9	28	26
Pelorus Jack	**J8x32 3C/4**	3.0	25	24
Montgomeries' Rant, The	**R8x32 3C/4**	2.5	22	21
Reel of the Royal Scots, The	**R8x32 3C/4**	2.3	27	20
Dream Catcher, The	**S96 Sq.**	3.2	19	19
Minister on the Loch, The	**S3x32 3C**	2.8	24	19
Irish Rover, The	**R8x32 3C/4**	3.9	20	18
Scott Meikle	**R4x32 4C**	3.4	17	17
Joie de Vivre	**J8x32 3C/4**	2.5	15	17
Maxwell's Rant	**R8x32 3C/4**	2.3	12	17
Belle of Bon Accord, The	**S4x32 4C**	3.4	16	16
Miss Johnstone of Ardrossan	**R5x32 5C**	3.1	15	16
Hooper's Jig	**J8x32 3C/4**	2.2	19	16
MacDonald of the Isles	**S3x32 3C**	2.8	16	15
Ian Powrie's Farewell to Auchterarder	**J128 Sq.**	3.9	13	15
Napier's Index	**J8x40 3C/4**	2.8	13	14
Reel of the 51st Division, The	**R8x32 3C/4**	2.2	14	14
Wind on Loch Fyne, The	**S3x32 Triang.**	3.0	16	14
Highland Rambler, The	**R8x40 3C/4**	2.5	14	14
Bratach Bana	**R8x32 3C/4**	3.7	13	14
Sailor, The	**R8x32 3C/4 (Hp)**	2.2	11	14
Anniversary Reel	**R4x32 4C**	3.7	14	13
Mrs Stewart's Jig	**J8x32 3C/4**	2.5	19	13

Using confidence levels

Assumes difficulty is the main difference between programmes

Set

Use variance to give weighted mean for programmes

$$p_i = \frac{\sum_{j \in P[i]} \frac{d_j}{\mathrm{Var}(d_j)}}{\sum_{j \in P[i]} \frac{1}{\mathrm{Var}(d_j)}} \qquad \mathrm{Var}(p_i) = \frac{1 + \sum_{j \in P[i]} \frac{(p_i - d_j)^2}{\mathrm{Var}(d_j)}}{\sum_{j \in P[i]} \frac{1}{\mathrm{Var}(d_j)}}$$

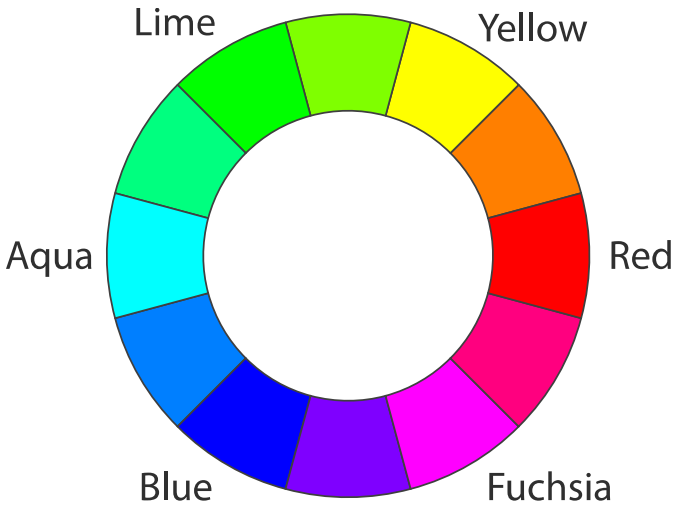
Then sort + normalize the results so the fall within 1 to 5 and follow a curve like people expect – a bit of heuristic.

Then iterate, do exactly the same to get dance difficulty from programme difficulty

Future

- Try out SVD – need to map from rankings for each group {Novice, beginner, Standard, Experienced, Demonstration} to overall difficulty ranking.

  Can one get a colour wheel angle from rankings for Red, Blue, Green for colour?



Other ideas

- Possibly have rank using compressibility of descriptions of the dances – database of over 4000 descriptions. Heuristic deciding size of dictionary formed by these.
- Popularity makes dances easier – cater for time component and area.
- Netflix overall winners improved on straight SVD with some heuristics, and other tweaks using search algorithms weighting the results with that of other algorithms.

Main sources + credits

- Matrix Factorization Techniques for Recommender Systems, byYehuda Korenn, Robert Bell and Chris Volinsky. 2009 Good introduction and where I got a couple of nice diagrams from
- From Tapestry to SVD: A Survey of the Algorithms That Power Recommender Systems. Joseph Huttner 2009
- A Singularly Valuable Decomposition: The SVD of a Matrix. Dan Kalman 2002
- Wikipedia: Recommender system, Singular value decomposition, Collaborative filtering. Also the source of the Google image and the logos
- What's Happening in the Mathematical Sciences 8: Accounting for Taste . Gabor Takacs. 2011